Residual Attribute Attention Network for Face Image Super-Resolution

Jingwei Xin,[†] **Nannan Wang,**[‡*] **Xinbo Gao,**[†] **Jie Li**[†]
[†] State Key Laboratory of Integrated Services Networks,
School of Electronic Engineering, Xidian University, Xi'an 710071, China
[‡] State Key Laboratory of Integrated Services Networks,
School of Telecommunications Engineering, Xidian University, Xi'an 710071, China

Abstract

Facial prior knowledge based methods recently achieved great success on the task of face image super-resolution (SR). The combination of different type of facial knowledge could be leveraged for better super-resolving face images, *e.g.*, facial attribute information with texture and shape information. In this paper, we present a novel deep end-to-end network for face super resolution, named Residual Attribute Attention Network (RAAN), which realizes the efficient feature fusion of various types of facial information. Specifically, we construct a multi-block cascaded structure network with dense connection. Each block has three branches: Texture Prediction Network (TPN), Shape Generation Network (SGN) and Attribute Analysis Network (AAN). We divide the task of face image reconstruction into three steps: extracting the pixel level representation information from the input very low resolution (LR) image via TPN and SGN, extracting the semantic level representation information by AAN from the input, and finally combining the pixel level and semantic level information to recover the high resolution (HR) image. Experiments on benchmark database illustrate that RAAN significantly outperforms state-of-the-arts for very low-resolution face SR problem, both quantitatively and qualitatively.

1. Introduction

Face super-resolution (SR), also known as face hallucination, is the process of recovering a high-resolution (HR) face from an input low-resolution (LR) face image. Face SR can be used as an important means of image preprocessing and widely used in various fields related to face images, *e.g.*, face parsing (Li 2017), identity recognition (Taigman 2014; Wang, Ye, and Yang 2018; Wang, Hu, and Yu 2016), and face alignment (Jourabloo, Ye, and Liu 2017), where high-frequency face details are desired.

Deep convolutional neural network (CNN) based SR methods have achieved significant improvements over conventional SR methods. Dong *et al.* (Dong, Loy, and He 2016) proposed SRCNN by firstly introducing CNN to image SR, which has established a nonlinear mapping from LR to HR image. Considering face hallucination is a domain-specific super-resolution problem, the prior knowledge in face images could be pivotal for face image super-resolution. Tuzel *et al.* (Tuzel, Taguchi, and Hershey 2016) proposed GLN to extract the global and local information from face images. Yu *et al.* (Yu and Porikli 2016) investigated GAN (Goodfellow, Pouget-Abadie, and Mirza 2014) to create perceptually realistic HR face images. Zhu *et al.* (Zhu, Liu, and Chen 2016) proposed CBN to overcome the different face spatial configuration by dense correspondence field estimation. Tai *et al.* (Tai, Chen, and Liu 2018) employed facial landmarks and parsing maps to train the network. Yu *et al.* (Yu, Basura, and Richard 2018) introduced a facial attribute embedding method into face image SR problem.

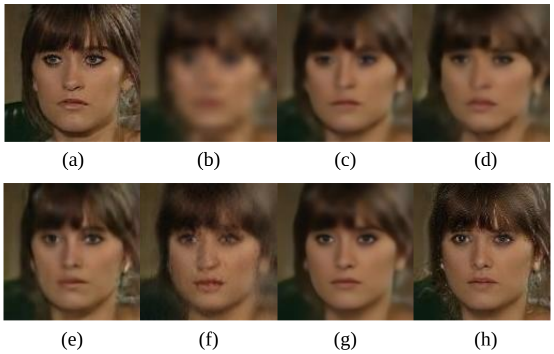

Figure 1: Face super-resolution results by state-of-the-art methods on scale factor 8. (a) Original HR images. (b) Input LR images. (c) Results of Ledig $et\ al.'s$ method (SRResNet) (Ledig, Theis, and Huszar 2017). (d) Results of Tuzel $et\ al.'s$ method (GLN) (Tuzel, Taguchi, and Hershey 2016). (e) Results of Huang $et\ al.'s$ method (Wavelet-SRNet) (Huang, He, and Sun 2017). (f) Results of Yu $et\ al.'s$ method (AEUN) (Yu, Basura, and Richard 2018). (g) Results of our RAAN. (h) Results of our RAAN-GAN.

Among them, the face prior information can roughly divided into two levels: the pixel level representation and the semantic level representation. Pixel level representation refers to the information of landmark, component and texture. Semantic level representation could be regarded as face attributes, *e.g.* age, gender and smile. However, most of existing methods explore only single type of face prior information, where prior information is not fully utilized. Fig. 1

[*] Corresponding author: Nannan Wang (nnwang@xidian.edu.cn)

presents the hallucinated details generated by state-of-the-art face SR methods. Neither the pixel level representation based methods SRResNet (Ledig, Theis, and Huszar 2017), GLN (Tuzel, Taguchi, and Hershey 2016),Wavelet (Huang, He, and Sun 2017) nor the semantic level representation based method AEUN (Yu, Basura, and Richard 2018) could learn the complex nonlinear mapping from LR to HR using limited facial information.

On the other hand, many previous face SR methods are complicated and difficult for application in real-world scenes. For instance, URDGN (Yu and Porikli 2016) and Attention-FH (Cao, Lin, and Shi 2017) requires that the input face images need to be pre-aligned. The Structured-FH (Yang, Liu, and Yang 2013) method and the CBN (Zhu, Liu, and Chen 2016) method adopt multi-stage network to reconstruction face image, rather than an end-to-end manner.

To practically resolve these problems, we propose a novel end-to-end Residual Attribute Attention Network (RAAN) for an easy trainable while effectively combining multi-level prior information. In this work, we divide the pixel level information of face images into texture feature and shape feature information. These features are taken as the basic features. The semantic level features are treated as a kind of control features. These control features could be used to guide the recombination of basic features to enhance the diversity of facial information.

Our RAAN is a multi-block cascaded network with dense connection, as shown in Fig.2. For each block, we first extract a set of coarse feature maps from the input, and feed them into three branches: Shape Generation Network (SGN), Texture Prediction Network (TPN) and Attribute Analysis Network (AAN). SGN is to generate facial global features such as shape and contours, TPN complements other information such as textures and components, and AAN extracts the attributes from the input image and guides the recombination of pixel level information by channel attention.

The main contributions of this work are threefold. (1) To the best of our knowledge, this is the first face super-resolution network which jointly utilizes facial prior information at pixel level and semantic level. (2) We first introduce the channel attention method into the face super resolution, and proposed a residual channel attention based multi-level information fusion strategy. (3) We proposed multi-dense connection structure to construct very deep trainable networks. The interstage connections and original connections could help the very deep network easy to train and avoid the problem of local convergence.

2. Related Work

Example-based SR methods achieved state-of-the-art performance by learning a mapping from LR image patches to HR image patches, *e.g.* dictionary learning-based methods (Lu, Yuan, and Yan 2012), local linear regression-based methods (Timofte, De, and Van Gool 2014), random forest-based method (Schulter, Leistner, and Bischof 2015) etc.. Recently, due to the outstanding learning ability, deep learning-based methods have demonstrated high superiority over classical example-based methods. Some works (Dong, Loy, and He 2016; Shi, Caballero, and Huszar 2016) have been introduce to the SR problem due to their excellent ability to learn knowledge from large scale of image patches. However, Because of the shallower network layers, theses networks have limited representation ability to fit the complicated nonlinear mapping from LR to HR. The very deep network (Kim, Lee, and Lee 2016; Ledig, Theis, and Huszar 2017; Tai, Yang, and Liu 2017) have been proposed to overcome this problem. A variety of skip connections are used in these methods, which leads these networks easily going deeper to enhance their representation ability. This also motivates us to employ a multi-block dense connection cascaded structure as our main network struture.

For face super-resolution, there are abundant facial prior information which could be used into the image reconstruction process. GLN transforms the input image into global and local feature maps by convolution and full connection. Zhu *et al.* (Zhu, Liu, and Chen 2016) presented an unified framework for face super-resolution and dense correspondence field estimation to recover textural details. They achieve state-of-the-art results for very low resolution inputs but fail on faces with various poses and occlusions due to the difficulty of accurate spatial prediction. Yu *et al.* (Yu and Porikli 2016) used the discriminant network with strong facial prior information to generate perceptually realistic HR face images. They further proposed transformative discriminative autoencoder to super-resolve unaligned, noisy and tiny LR face images (Yu and Porikli 2017). Cao *et al.* (Cao, Lin, and Shi 2017) proposed an attention-aware face hallucination framework, which resorts to deep reinforcement learning for sequentially discovering attended patches and then performs the facial part enhancement by fully exploiting the global image interdependency. Huang *et al.* (Huang, He, and Sun 2017) proposed an Wavelet-based CNN method, which learns to predict the LR's corresponding series of HR's wavelet coefficients, and utilizes them to reconstructing HR images. Chen *et al.* (Tai, Chen, and Liu 2018) introduced facial landmarks and parsing maps to train the network by multi-supervision. Yu *et al.* (Yu, Basura, and Richard 2018) proposed an attribute embedding based coding and decoding network, which first encodes LR images with facial attributes and then super-resolves the encoded features to hallucinate LR face images.

Attention has been introduced by many recent works as a method to enhance the ability of feature extraction. Hu *et al.* (Hu, Shen, and Sun 2017) focuses on channels and proposed Squeeze-and-Excitation block, which adaptively recalibrates channel-wise feature responses by modeling the interdependencies between channels. Wang *et al.* (Wang, Jiang, and Qian 2017) proposed residual attention network with a trunk-and-mask attention mechanism to obtain significant performance improvement. Attention is actually a recombination of the information flow in the network. Its purpose is to increase the proportion of information-rich regions and restrain redundant information to improve the nonlinear mapping ability of the network. However, existing attention vectors are used only as the importance of feature maps, but lack further definitions. In this work, We redefine the chan-

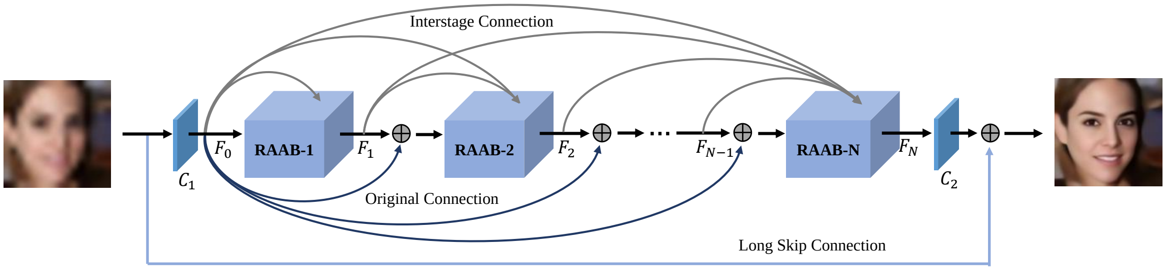

Figure 2: Pipeline of our proposed RAAN model.

nel attention with the facial attributes, and introduce it to the task of face image SR. Experimental results proved the significance of our attribute attention.

3. Residual Attribute Attention Network

The pipeline of our proposed RAAN model is shown in Fig. 2. Let x denote the LR input image and y as the recovered high-resolution image. We first extract the feature maps from the input LR image, and regard them as the basic features as follows,

$$F_0 = C_1(x), \tag{1}$$

where C_1 denotes the mapping from a LR image x to basic feature maps F_0 implemented by some convolution layers. Then, F_0 is fed into the network of multiple stacked residual attribute attention blocks. For each block, we have

$$F_n = R_n(F_{n-1}; F_{n-1}, F_{n-2}, ..., F_0), \tag{2}$$

where F_n is the output of the n_{th} Residual Attribute Attention Block (RAAB), R_n is the block function and N is the number of RAAB. The input of each block can be divided into two parts: F_{n-1}, which is the direct signal input between adjacent blocks, and $F_{n-1}, F_{n-2}, ..., F_0$ which are the output of previous blocks and are transmitted to current block by Interstage Connection. To the best of our knowledge, our multiple RAAB achieves the largest depth so far and provides very large receptive field size. So we treat its output as final residual features, and directly acquire HR face image by C_2 as follows,

$$y = C_2(F_N) + x. \tag{3}$$

Given a training set $\{x^{(i)}, \widehat{att}^{(i)}, \widehat{y}^{(i)}\}_{i=1}^{M}$, M is the number of training images, $\widehat{y}^{(i)}$ is the ground-truth HR image corresponding to the LR image $x^{(i)}$, and $\widehat{att}^{(i)}$ is the corresponding ground-truth attribute information. The loss function of our RAAN is

$$L_G(\theta) = \frac{1}{M}\sum_{i=1}^{M}\{\left\|\widehat{y}^{(i)} - y^{(i)}\right\| + \lambda\sum_{n=1}^{N}\left\|\widehat{att}^{(i)}_{(n)} - att^{(i)}_{(n)}\right\|\}, \tag{4}$$

where θ denotes the parameters, λ is the trade-off between the attribute prior information and the prediction loss, $y^{(i)}$ and $att^{(i)}_{(n)}$ are the recovered HR image and the estimated prior attributes for the i_{th} image at the n_{th} RAAB block.

3.1 Details on RAAB

For each set of basic image features F_n, we fed it into the Residual Attribute Attention Block for more accurate facial feature information. In this part, we present the details of our RAAB, as shown in Fig. 3. It contains one feature extraction block and three branches. Different type of information are fused at the end of the block.

Feature Extraction Block Considering different sensitivities of three branches to each type of information, we first pre-extract the features from the input feature maps by a Feature Extract Block (FEB), aiming to obtain the desired features easier for next three branches. The architecture of the FEB is shown in the Fig. 3, which consists of four 3 × 3 convolution layers:

$$F_{n,P} = P(F_n), \tag{5}$$

where $P()$ and $F_{n,P}$ is the function and output of the FEB respectively.

SGN and TPN The pixel level information of face image could be divided into shape and texture information. Considering the shape information is better preserved compared to the texture when reducing the resolution from high to low, we propose the SGN structure to generate the shape information, which consists of an encoder and a decoder. The encoder continuously down-samples the image to remove the texture information. The decoder then recovers the shape features to the same size as the input image. In addition, we establish many skip connections between the encoder and the decoder. Our structure of shape generation network is shown in Fig. 3. For the pre-extract features $F_{n,P}$, we have

$$F_{n,P,S} = S_2(S_1(F_{n,P})), \tag{6}$$

where $S_1()$ and $S_2()$ are the encoding and decoding model respectively, and $F_{n,P,S}$ is shape features of the input image. In terms of the texture information, we propose TPN to supplement the texture features in the network. It is well known that the receptive field of the network has a remarkable effect on the ability of nonlinear mapping. A larger receptivity field can make the network more sensitive for the details of

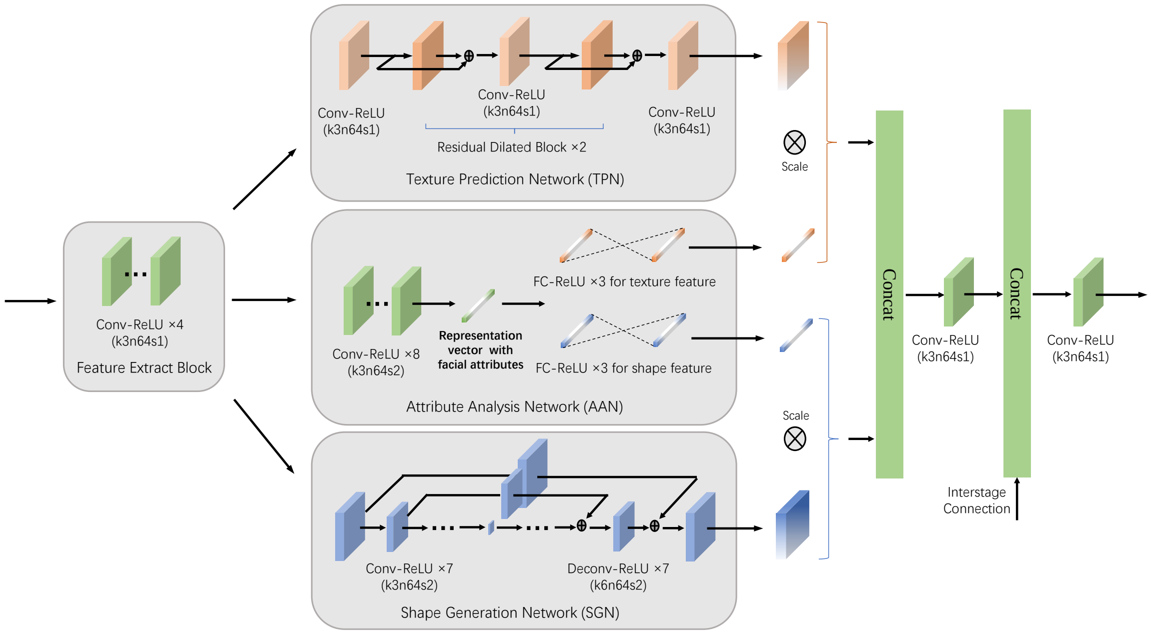

Figure 3: Structure of Residual Attribute Attention Block (RAAB).

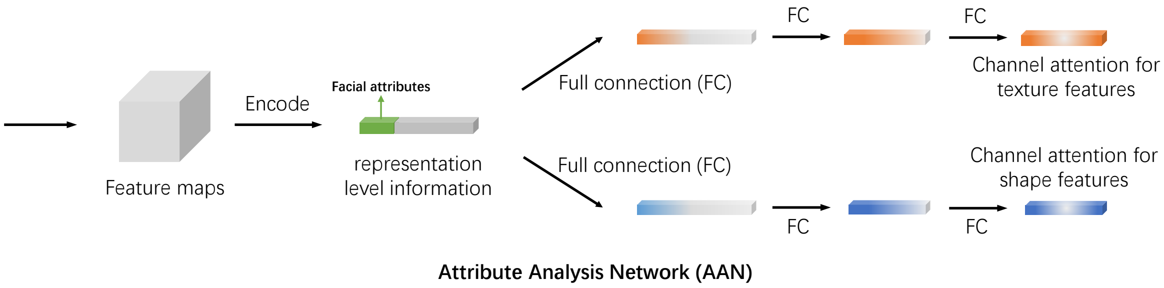
Feature maps
Encode
Facial attributes
representation level information
Full connection (FC)
FC
FC
Channel attention for texture features
Full connection (FC)
Channel attention for shape features
FC
FC
Attribute Analysis Network (AAN)

Figure 4: Detail Structure of Attribute Attention Network (AAN).

the input signal, and facilitate the extraction of texture information. Thus our TPN starts with a 3 × 3 convolution layer, and followed by two dilated blocks, and then another 3 × 3 convolution layer is used to reconstruct the texture graphs. The dilated convolution can effectively increase the mapping range of each pixel between adjacent network layers, and improve the receptivity field of each layer of the network. For each dilated block, we have three dilated convolution layers with dilate coefficient 2, 3 and 2 respectively. Our model of TPN could be formulated as

$$F_{n,P,T} = T_3(D_2(T_2(D_1(T_1(F_{n,P}))))), \tag{7}$$

where T_1, T_2, T_3 are three 3×3 convolution layers, D_1 and D_2 represent two dilated blocks, and $F_{n,P,T}$ is the texture features of the input image. SGN and TPN are employed to extract pixel level information. Then, we make further enhancements to these pixel information as introduced in the following subsection.

Attribute Attention Network Compared with natural images, face images have more prior information and can be used to facilitate image reconstruction, such as face attributes. Previous works (Tuzel, Taguchi, and Hershey 2016; Ledig, Theis, and Huszar 2017; Huang, He, and Sun 2017) only take LR images as inputs and then super-resolve them by convolutional layers, or only introduce the single type of prior information to the network (Yu, Basura, and Richard 2018; Tai, Chen, and Liu 2018). These works have not taken the correlation among diverse prior information into account.

In this part, We introduce the attribute information of semantic level representation on the basic of pixel level features. We use the attributes to define the distribution of face pixel level information by the means of channel attention. Considering the information passing through SGN and TPN is inevitably lost because of their different tendencies for feature extraction, the input of AAN is consistent with SGN and TPN. We first acquire the high-level features of the input image by an encoder which stacks multiple convolution layers. As shown in Fig. 4, for the set of high-level representation information, we selected one of the sections as the attribute information, and it will be supervised during the training of the network. Then, this set of high-level features is fed into the two full connection blocks for the channel attention. The outputs of two full connection blocks are corresponding to the shape feature and the texture feature respectively. AAN could be formulated as,

$$A_{n,a} = A_t(F_{n,P}), \tag{8}$$

$$R_{n,a,t} = F_t(A_{n,a}), \tag{9}$$

$$R_{n,a,s} = F_s(A_{n,a}), \tag{10}$$

where $A_t()$ is the function of attribute encoder, $A_{n,a}$ is the representation level information with facial attributes, $F_t()$ and $F_s()$ are two full connection blocks, and which output $R_{n,a,t}$ and $R_{n,a,s}$ are the residual attribute attention of the texture and shape features. Those are then multiplied to the output of TPN and SGN, for further enhancing the pixel level information (Hong, Yu, and Wan 2015).

Multi-dense connection structure We proposed a multi-dense connection structure, which can be divide it into interstage connection and original connection. The commonly used dense connection back propagates the gradient of the network more easily during the training process (Szegedy, Vanhoucke, and Ioffe 2016; Huang, Liu, and Weinberger 2017). However, in this work, very low resolution face images might map to many high resolution face candidates during the process of making them high resolution. Because the network lacks the ability to distinguish the features, it only trains in the direction of decreasing the loss gradient quickly. So, the classical dense connection would make the network more prone to local convergence and to get a low performace face image. To avoid this phenomenon, we proposed the multi-dense connection. At first, the interstage connection could make the network gradient more easily back propagation, and extend the range of perception for each block of feature maps. Then, the original connection improves the network utilization of input image information by enhancing the original information between the adjacent blocks of the network, and avoids the phenomenon of local convergence.

3.2 RAAN-GAN

GAN-based methods achieve good visual effect for image synthesis (Ledig, Theis, and Huszar 2017). Because of its prominent features (such as symmetry of contour, similarity of components), we propose to incorporate GAN into our framework. The key idea is to use a discriminant network to distinguish the super-resolved images and the real high-resolution images, and also to train the SR network to deceive the discriminator.

Our discriminant network consists of eight convolution layers and two full connection layers. The objective function of the adversarial network D is expressed as

$$\begin{aligned} L_D(G, D) = \mathrm{E}[\log D(\widehat{y}, x)] \\ +\mathrm{E}[\log(1 - D(G(x), x))], \end{aligned} \tag{11}$$

where E is the expectation of the log probability distribution and D is the generative model. The loss function of our generative model could be formulated

$$\arg \min_{G} \max_{D} L_G(\theta) + \gamma L_D(G, D) \tag{12}$$

where γ is the trade-off between the discriminant loss and the aforementioned RAAN loss.

4. Experimental Results and Analysis

We conduct extensive experiments on celebA dataset (Liu, Luo, and Wang 2015). We use the first 18000 images for training, and the following 100 images for testing. We coarsely crop the training images according to their face regions and resize to 128×128 without any pre-alignment operation. Here we use color images for training as SRGAN does (Ledig, Theis, and Huszar 2017). The input low-resolution images are firstly resized by bicubic interpolation to the same size as the output high-resolution images. We implement our moel using the pytorch environment. Adam with an initial learning rate of 3×10^{-4} are used in our model. The batch size is set to 16. We empirically set $\lambda = 1$ and $\gamma = 0.01$. Training a basic RAAN on celebA dataset generally takes 5 hours with one Titan X Pascal GPU.

4.1 Comparisons with State-of-the-Art Methods

We compare our proposed RAAN and RAAN-GAN with state-of-the-art SR methods, including generic SR methods like SRResNet (Ledig, Theis, and Huszar 2017), VDSR (Kim, Lee, and Lee 2016) and SRCNN (Dong, Loy, and He 2016); and facial SR methods like GLN (Tuzel, Taguchi, and Hershey 2016), Wavelet-SRNet (Huang, He, and Sun 2017) and FSRNet (Tai, Chen, and Liu 2018). Aiming at fair comparison, we train all models (except the FSRNet) with the same training set. For FSRNet, We use the trained model provided by the authors to directly generate results. FSRNet choose first 18000 images for training, and rotate images by 90°, 180°, 270° and flip them horizontally, which training dataset is bigger than ours.

we compare RAAN with the state-of-the-arts quantitatively. Tab. 1 summarizes quantitative results on the Celeba datasets. Our RAAN significantly outperforms state-of-the-arts in both PSNR and SSIM. Qualitative comparisons of RAAN/RAAN-GAN with prior works are illustrated in Fig. 5. We follow the same experimental setting on handling occlued face as Wavelet-SRNet (Huang, He, and Sun 2017) and directly import the 16×16 test examples for super-resolving 128×128 HR images. Benefiting from the facial attribute attentions, our method produces relatively sharper edges and shapes, while other methods may give more blurry results. Moreover, RAAN-GAN further recovers sharper facial textures than RAAN.

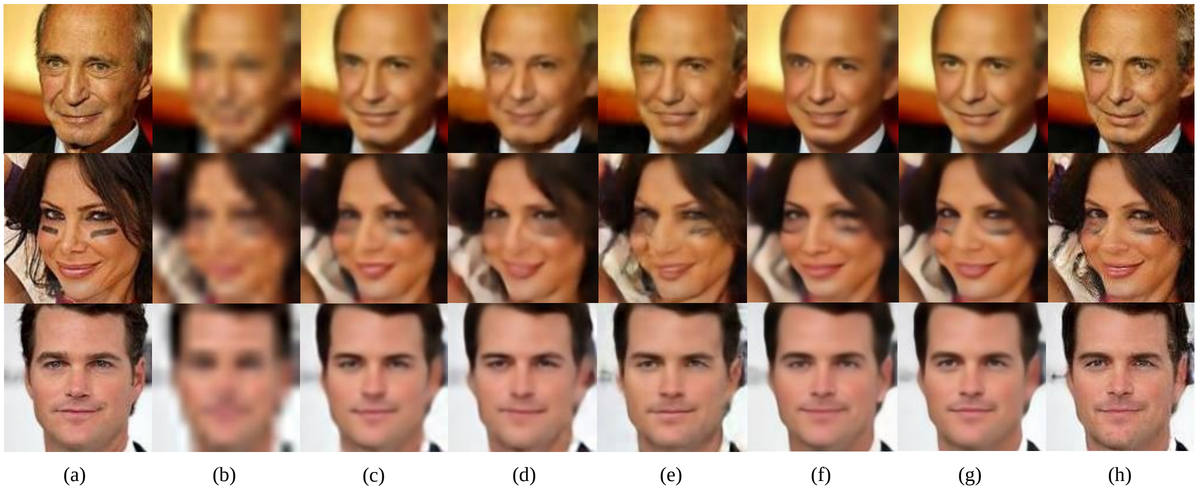

Figure 5: Face super-resolution examples by state-of-the-art methods on scale 8. (a) Original HR images. (b) Input LR images. (c) Results of Ledig $et\ al.'s$ method (SRResNet) (Ledig, Theis, and Huszar 2017). (d) Results of Tuzel $et\ al.'s$ method (GLN) (Tuzel, Taguchi, and Hershey 2016). (e) Results of Huang $et\ al.'s$ method (Wavelet-SRNet) (Huang, He, and Sun 2017). (f) Results of Tai $et\ al.'s$ method (FSRNet) (Tai, Chen, and Liu 2018). (g) Results of our RAAN. (h) Results of our RAAN-GAN.

Scale	Bicubic PSNR/SSIM	SRCNN PSNR/SSIM	VDSR PSNR/SSIM	SRResNet PSNR/SSIM	GLN PSNR/SSIM	Wavelet-SRNet PSNR/SSIM	FSRNet PSNR/SSIM	RAAN PSNR/SSIM
×4	28.07/0.8097	29.55/0.8456	30.36/0.8632	30.48/0.0.8674	**30.60/0.8685**	29.97/0.8612	-	**31.22/0.8804**
×8	24.06/0.6535	24.98/0.6924	26.17/0.7410	26.48/0.7529	25.94/0.7371	26.16/0.7460	**26.82/0.7601**	**27.03/0.7740**

Table 1: Benchmark super-resolution, with PSNR/SSIMs for scale 8. The **red** and **blue** text indicates the best/second best performance.

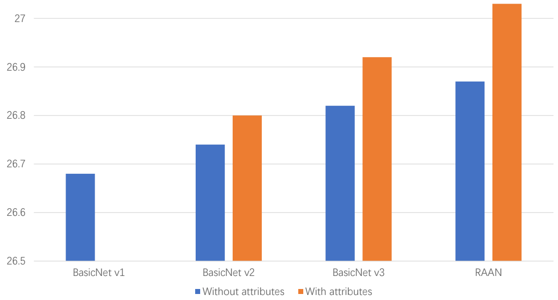

Figure 6: Ablation study on effects of different attribute attentions.

4.2 Model Analysis

We conduct ablation study on the effects of attribute attention, and clearly show how the performance improves with different kinds of feature attention. In this test, we estimate the facial attributes through the attribute analysis network instead of using the ground truth conducted. Same as the tests conducted in Fig. 6, we conduct 7 experiments to estimate the different kinds of features. Specifically, by removing the attribute analysis network from our basic RAAN, the remaining parts constitute the first network, named ‘BasicNet v1’, which has a dual network structure. The second network named ‘BasicNet v2’, and The third network named ‘BasicNet v3’, them based on the ‘BasicNet v1’ have the attribute analysis network but just connected to the SGN and TPN respectively.

Fig. 6 shows the results of different network structures. It can be seen that: 1.Compared to the first networks, the supervision on facial attribute further improves the performance, which indicates the estimated facial priors indeed have positive effects on face super-resolution. 2. The enhancement of facial texture information, could more obvious promote the performance of face image super-resolution than the facial shape information. 3. The model using both attentions achieves the best performance, which indicates richer facial information brings more improvement.

Then, we also focus on the different kinds of skip connections, and conduct 4 experiments. As shown in Fig. 8, ‘Basic’ is our network but without skip connection, ‘IC’ and ‘OC’ corresponding to the Interstage Connection (IC) and Original Connection (OC). It can be find that OC could promote the performance of network but IC not. The combination of IC and OC (RAAN) could get better performance.

4.3 Subjective Visual Evaluation

In this section, we compaired our RAAN and RAAN-GAN with the state-of-the-art GAN based methods like URDGN, TDAE, AEUN and FSRGAN. Except the FSRGAN, we also train these models with the same training set. As shown in

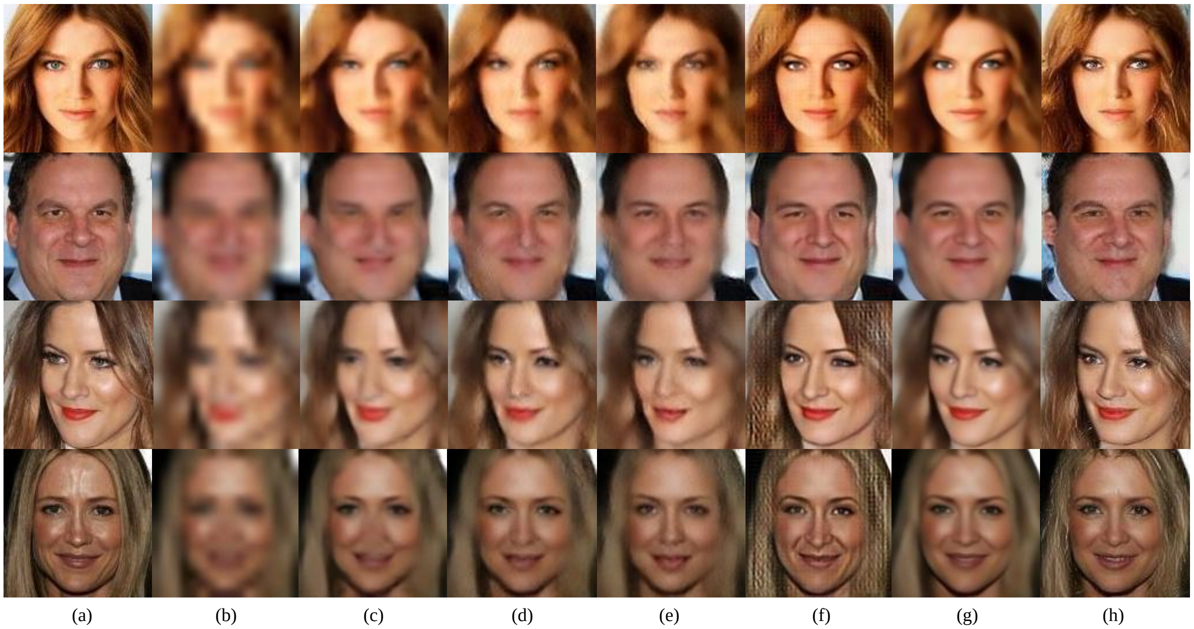

Figure 7: Visual evaluation on scale 8. (a) Original HR images. (b) Input LR images. (c) Results of Yu $et\ al.'s$ method (URDGN) (Yu and Porikli 2016). (d) Results of Yu $et\ al.'s$ method (TDAE) (Yu and Porikli 2017). (e) Results of Yu $et\ al.'s$ method (AEUN) (Yu, Basura, and Richard 2018). (f) Results of Tai $et\ al.'s$ method (FSRGAN) (Tai, Chen, and Liu 2018). (g) Results of our RAAN. (h) Results of our RAAN-GAN.

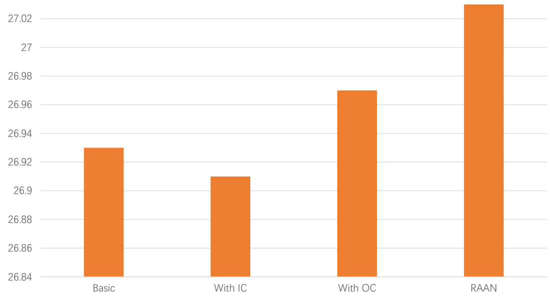

Figure 8: Ablation study on effects of skip connection.

Fig. 7, Our RAAN-GAN achieves better visual effects than the state-of-the-arts, especiall the texture of hair and face. The structure of URDGN is relatively simple, the effect of model mapping is slightly inadequate. AEUN can be seen as an improvement by introduction the face attribute information to TADE, thus the individual components of face image will be generated more clearly. FSRGAN can generate face images with clear contour by their Prior Estimation Network, but slight shortage of texture generation.

Conclusion

In this paper, a novel deep end-to-end trainable Residual Attribute Attention Network (RAAN) is proposed for face super-resolution. The key contribute of RAAN is the method of residual attribute attention. We divide the information of face image into the shape features and texture features. And according to the attribute information which learned from the input LR face image, achieving the recombination of these two kinds of information by channel attention. Extensive benchmark experiments show that RAAN significantly outperforms state-of-the-arts. In addition, we also proposed a multi-dense connection method, It could make the very deep network easy training and avoid the problem of local convergence.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61876142, 61432014, U1605252, 61671339, 61772402, 61501339, and in part by the National Key Research and Development Program of China under Grant 2016QY01W0200, in part by National High-Level Talents Special Support Program of China under Grant CS31117200001, in part by Young Elite Scientists Sponsorship Program by CAST (under Grant 2016QNRC001), in part by Natural Science Basic Research Plan in Shaanxi Province of China (under Grant 2017JM6085 and 2017JQ6007), in part by Young Talent fund of University Association for Science and Technology in Shaanxi, China, in part by the Fundamental Research Funds for the Central Universities under Grant XJS17086, in part by CCF-Tencent Open Fund, in part by the Xidian University-Intellifusion Joint Innovation Laboratory of Artificial Intelligence, and in part by the Innovation Fund of Xidian University.

References

Cao, Q.; Lin, L.; and Shi, Y. 2017. Attention-aware face hallucination via deep reinforcement learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1656–1664. IEEE.

Dong, C.; Loy, C.; and He, K. 2016. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(2):295–307.

Goodfellow, I.; Pouget-Abadie, J.; and Mirza, M. 2014. Generative adversarial nets. In *International Conference on Neural Information Processing Systems*, 2672–2680. MIT Press.

Hong, C.; Yu, J.; and Wan, J. 2015. Multimodal deep autoencoder for human pose recovery. *IEEE Transactions on Image Processing* 24(12):5659–5670.

Hu, J.; Shen, L.; and Sun, G. 2017. Squeeze-and-excitation networks. In *arXiv*. arXiv.

Huang, H.; He, R.; and Sun, Z. 2017. Wavelet-srnet : A wavelet-based cnn for multi-scale face super resolution. In *IEEE International Conference on Computer Vision*, 1698–1706. IEEE.

Huang, G.; Liu, Z.; and Weinberger, K. 2017. Densely connected convolutional networks. In *IEEE conference on computer vision and pattern recognition*. IEEE.

Jourabloo, A.; Ye, M.; and Liu, X. 2017. Pose-invariant face alignment with a single cnn. In *IEEE International Conference on Computer Vision*, 3219–3228. IEEE.

Kim, J.; Lee, J.; and Lee, K. 2016. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1646–1654. IEEE.

Ledig, C.; Theis, L.; and Huszar, F. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 105–114. IEEE.

Li, Y. Liu, S. Y. J. 2017. Generative face completion. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Liu, Z.; Luo, P.; and Wang, X. 2015. . deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, 3730–3738. IEEE.

Lu, X.; Yuan, H.; and Yan, P. 2012. Geometry constrained sparse coding for single image super-resolution. In *Computer Vision and Pattern Recognition*, 1648–1655. IEEE.

Schulter, S.; Leistner, C.; and Bischof, H. 2015. Fast and accurate image upscaling with super-resolution forests. In *Computer Vision and Pattern Recognition*, 3791–3799. IEEE.

Shi, W.; Caballero, J.; and Huszar, F. 2016. Real-time single image and video super-resolution using an efficient subpixel convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1874–1883. IEEE.

Szegedy, C.; Vanhoucke, V.; and Ioffe, S. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. IEEE.

Tai, Y.; Chen, Y.; and Liu, X. 2018. Fsrnet:end-to-end learning face super-resolution with facial priors. In *IEEE conference on computer vision and pattern recognition*. IEEE.

Tai, Y.; Yang, J.; and Liu, X. 2017. Image super-resolution via deep recursive residual network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2790–2798. IEEE.

Taigman, Y. Yang, M. R. M. 2014. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1701–1708. IEEE.

Timofte, R.; De, V.; and Van Gool, L. 2014. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian Conference on Computer Vision*, 111–126.

Tuzel, O.; Taguchi, Y.; and Hershey, J. 2016. Global-local face upsampling network. In *arXiv*. arXiv.

Wang, Z.; Hu, R.; and Yu, Y. 2016. Scale-adaptive low-resolution person re-identification via learning a discriminating surface. In *International Joint Conference on Artificial Intelligence*, 2669–2675.

Wang, F.; Jiang, M.; and Qian, C. 2017. Residual attention network for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6450–6458. IEEE.

Wang, Z.; Ye, M.; and Yang, F. 2018. Cascaded sr-gan for scale-adaptive low resolution person re-identification. In *International Joint Conference on Artificial Intelligence*, 3891–3897.

Yang, C.; Liu, S.; and Yang, M. 2013. Structured face hallucination. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1099–1106. IEEE.

Yu, X., and Porikli, F. 2016. Ultra-resolving face images by discriminative generative networks. In *European Conference on Computer Vision*, 318–333. Springer.

Yu, X., and Porikli, F. 2017. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5367–5375. IEEE.

Yu, X.; Basura, F.; and Richard, H. 2018. Super-resolving very low-resolution face images with supplementary attributes. In *IEEE conference on computer vision and pattern recognition*. IEEE.

Zhu, S.; Liu, S.; and Chen, C. 2016. Deep cascaded bi-network for face hallucination. In *European Conference on Computer Vision*, 614–630. Springer.